

ICS 35.240.01

**Künstliche Intelligenz –
Life Cycle Prozesse und Qualitätsanforderungen –
Teil 3: Erklärbarkeit; Text Englisch**

Artificial Intelligence –
Life Cycle Processes and Quality Requirements –
Part 3: Explainability; Text in English

Intelligence artificielle –
Processus du cycle de vie et exigences pour la qualité –
Partie 3: Explicabilité; Texte en anglaise

Gesamtumfang 24 Seiten

Dieses Dokument wurde durch die im Vorwort genannten Verfasser erarbeitet und verabschiedet.

Contents

	Page
Foreword	4
Introduction	6
1 Scope	7
2 Normative references	7
3 Terms and definitions	7
4 Life Cycle, Stakeholders, and Information Needs	8
4.1 General	8
4.2 AI System Life Cycle	9
4.3 Stakeholder Roles	9
4.4 Interactions & Information Needs	9
4.4.1 General	9
4.4.2 Develop	11
4.4.3 Provide	11
4.4.4 Use	11
4.4.5 Evaluate	12
5 Opacity	12
5.1 General	12
5.2 What is Opacity?	12
5.3 Sources of Opacity	12
5.3.1 General	12
5.3.2 Technical Illiteracy	13
5.3.3 Intentional Obfuscation	13
5.3.4 System Complexity	13
5.4 Effects of Opacity	13
5.4.1 Effects on Development	13
5.4.2 Effects on Provision	14
5.4.3 Effects on Use	14
5.4.4 Effects on Evaluation	14
6 Explainability	15
6.1 General	15
6.2 Explanation & Explainability	15
6.3 Explanans: Content and Form	15
6.4 Interpretation & Interpretability	16
6.5 When is explanation needed?	16
6.5.1 General	16
6.5.2 Reliability	16
6.5.3 Stakes	16
6.6 Risks of Explanation	17
7 Providing Explanations	17
7.1 General	17
7.2 Classification of Explanation Methods	17
7.3 Explanation Qualities	18
7.3.1 Explanation documentation	18
7.3.2 Fidelity and accuracy	18
7.3.3 Stability and consistency	19
7.3.4 Psychological appropriateness	19
7.3.5 Conceptual Alignment	19
7.3.6 Explanation Variety	20
7.3.7 Completeness	20
8 Explainability by Design	20
8.1 General	20

8.2 Requirement Analysis 21

8.3 Explanation Technical Design 21

8.4 Explanation Validation 21

Bibliography 22

Figures

Figure 1 — The Explainability by Design process 20

Tables

Table 1 — Stakeholder roles (columns) and life cycle stages (rows) 10

Foreword

This DIN SPEC has been developed according to the PAS procedure. The development of a DIN SPEC according to the PAS procedure is carried out in DIN SPEC (PAS)-consortiums and does not require the participation of all stakeholders.

This DIN SPEC (PAS) is the result of „Zertifizierung von Standard-KI-Anwendungen (Zertifizierte KI)“ [en: „Certification of Standard-AI-Applications (Certified AI)“] funded by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia (funding code 005-2011-0053).

This document has been developed and adopted by the initiator(s) and authors named below:

- ai.dopt GmbH
André Bluhm
- Cleopa GmbH
Detlef Olschewski
- DEKRA Digital GmbH
Dr. Jörg Firnkorn
- Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Adriano Lucieri, Dr. Sebastian Palacio
- Deutsches Zentrum für Schienenverkehrsforschung beim Eisenbahn-Bundesamt (DZSF)
Dr. Rustam Tagiew
- eco – Verband der Internetwirtschaft e. V.
Hauke Timmermann
- Eindhoven University of Technology (TU/e)
Yeji Streppel, Dr. Carlos Zednik
- Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Rebekka Görge, Dr. Maximilian Poretschkin
- Gesellschaft für Informatik e. V.
Nikolas Becker
- Hochschule Düsseldorf
Prof. Dr. Thomas Zielke
- IAV GmbH

Dr. Christian Kruschel

— Leibniz-Fachhochschule

Prof. Dr. Matthias Neumann-Brosig

— metafinanz Informationssysteme GmbH

Stephen Bäuerle, Márton Eifert, Erik Martori López

— neurocat GmbH

Felix Assion, Annegrit Seyerlein-Klug

— Otto-Friedrich-Universität Bamberg

Prof. Dr. Ute Schmid

— Physikalisch-Technische Bundesanstalt (PTB)

Prof. Dr. Stefan Haufe

— QuantPi GmbH

Lukas Bieringer, Dr. Antoine Gautier

— TRB Holding UG (haftungsbeschränkt)

Dr. Tarek R. Besold

— Universität Bonn

Prof. Dr. Armin B. Cremers

— VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V.

Andreas Hauschke

At present, there are no standards covering this topic in the body of German Standards.

DIN SPEC (PAS)s are not part of the body of German Standards.

A draft of this DIN SPEC (PAS) has not been published.

Despite great efforts to ensure the accuracy, reliability and precision of technical and non-technical information, the DIN SPEC (PAS)-consortium cannot give any explicit or implicit assurance or warranty in respect of the accuracy of the document. Users of this document are hereby made aware that the consortium cannot be held liable for any damage or loss. The application of this DIN SPEC (PAS) does not release users from the responsibility for their own actions and is applied at their own risk.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. DIN shall not be held responsible for identifying any or all such patent rights.

Provision of this document free of charge as a PDF via the Beuth WebShop has been financed in advance.

For current information on this standard, please go to DIN's website (www.din.de) and search for the document number in question.

Introduction

The discipline of Artificial Intelligence (AI) aims to develop systems capable of performing tasks that would otherwise require human intelligence. Although the principles governing these systems' operation continues to be the subject of ongoing research, their increasingly high performance has led to their proliferation in many different application domains.

In order to ensure that AI systems are used efficiently, responsibly, and in a trustworthy way, their development and deployment should satisfy relevant quality criteria. Beyond the quality criteria for other kinds of computing systems as defined in ISO/IEC/IEEE 12207:2017, AI systems call for special-purpose criteria.

This is the third document in a series. DIN SPEC 92001-1 establishes an AI quality meta-model and life cycle to highlight the general AI quality criteria of robustness, explainability, and performance. DIN SPEC 92001-2 specifies criteria to ensure AI system robustness. This document, DIN SPEC 92001-3, specifies criteria to promote explainability.

1 Scope

This document provides a domain-independent guide of appropriate approaches and methods to promote explainability throughout all stages of the AI system life cycle. It defines 'opacity', describes the sources and effects of opacity in contemporary AI, and considers the way in which explanations can and should be employed to mitigate these effects for different stakeholders at different stages of the AI system life cycle.

The quality criteria described in this document aim to be comprehensive with respect to all kinds of AI systems, and with respect to all domains in which such systems are deployed. In particular, they apply to AI systems of different degrees of opacity. Thus, these quality criteria facilitate a better understanding and more effective use of explainable AI (XAI), allowing organizations to embed principles of trust and transparency into their AI-based applications. They also provide the conceptual and procedural basis for formulating specific requirements of explainability and comprehensible evidence of technically correct operation.

This document is not a requirements specification and does not provide any implementation requirements.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

- ISO Online browsing platform: available at <https://www.iso.org/obp>
- IEC Electropedia: available at <https://www.electropedia.org/>

DIN and DKE maintain terminological databases for use in standardization at the following addresses:

- DIN-TERMinologieportal: available at <https://www.din.de/go/din-term>
- DKE-IEV: available at <https://www.dke.de/DKE-IEV>

3.1

opacity

opaqueness

property of a system that appropriate information about the system is unavailable to relevant stakeholders

Note 1 to entry: Opacity is the opposite of transparency.

3.2

explanation

process of describing and communicating important factors influencing an AI system's behavior to a relevant stakeholder

Note 1 to entry: Explanation aims to reduce opacity (3.1).

Note 2 to entry: Explanation differs from interpretation (3.7).

3.3

explanans

product or output of an explanation

Note 1 to entry: An explanans can vary in both content and form.

3.4

explainability

property of an AI system to express important factors influencing the AI system results in a way that humans can understand

Note 1 to entry: In DIN SPEC 92001-1, the term 'comprehensibility' is used instead.

[SOURCE: ISO/IEC 22989:2022, 3.5.7, modified — added Note 1.]

3.5

explainability by design

design process within the software engineering workflow for ensuring an AI system's explainability

Note 1 to entry: Explanation by design spans the whole AI system life cycle.

3.6

explainable artificial intelligence

XAI

<discipline> research and development of methods to explain an AI system's behavior

3.7

interpretation

assignment of meaning to elements or features of an AI system

3.8

interpretability

property of an AI system that elements or features can be assigned meanings

Note 1 to entry: An AI system can be considered interpretable before or after the provision of explanations.

3.9

fairness

treatment, behaviour or outcomes that respect established facts, societal norms and beliefs and are not determined or affected by favouritism or unjust discrimination

Note 1 to entry: Considerations of fairness are highly contextual and vary across cultures, generations, geographies and political opinions.

Note 2 to entry: Fairness is not the same as the lack of bias. Bias does not always result in unfairness and unfairness can be caused by factors other than bias.

[SOURCE: ISO/IEC TR 24368:2022, 3.7]

4 Life Cycle, Stakeholders, and Information Needs

4.1 General

The quality criteria for explainability (7.3) promote different stakeholders' ability to perform their designated roles at different stages of the AI system life cycle. At each stage, a particular stakeholder has specific information needs, the satisfaction of which allows them to perform their designated roles and to interact with the AI system in the expected ways.